E-Commerce Letters 电子商务评论, 2024, 13(4), 2172-2179
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341380

Hans汉斯

基于LDA-SVM模型对企业债券的违约研究

陈　伟

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2024年8月29日；录用日期：2024年9月24日；发布日期：2024年11月8日

摘　要

近年来，债券市场已不再是稳固收益。随着第一支债券违约发生，债券市场打破刚性兑付情况。而债券违约预测成为众多学者关注重点。本文以发行债券的150家企业为研究对象，通过分析各企业一年期的财务指标，反应企业偿债能力、盈利能力等多个方面。以fisher线性判别分析(linear discriminant analysis, LDA)进行降维，构建了LDA-SVM模型对债券违约与否进行预测。并对比了逻辑回归(logit)、支持向量机(Support Vector Machine, SVM)、XGboost等二分类预测模型，其结果表明本文模型效果显著，有90%的准确率，对债券违约预测提供了有效思路，为投资人对于债券违约风险提供了参考。

关键词

债券违约，财务指标，线性判别分析，支持向量机

Research on the Default of Corporate Bonds Based on LDA-SVM Model

Wei Chen

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Aug. 29th, 2024; accepted: Sep. 24th, 2024; published: Nov. 8th, 2024

Abstract

In recent years, the bond market has ceased to be a solid yield. With the default of the first bond, the bond market broke the rigid payment situation. The prediction of bond default has become the focus of many scholars. This paper takes 150 companies that issued bonds as the research object, and analyzes the one-year financial indicators of each enterprise to reflect the solvency and profitability of enterprises. Fisher linear discriminant analysis (LDA) was used to reduce dimensionality, and the LDA-SVM model was constructed to predict whether the bond would default or not. The study compared several binary classification prediction models, including Logistic Regression (logit), Support Vector Machine (SVM), and XGBoost. The results showed that the model proposed in this

文章引用: 陈伟. 基于 LDA-SVM 模型对企业债券的违约研究[J]. 电子商务评论, 2024, 13(4): 2172-2179.
DOI: 10.12677/ecl.2024.1341380

paper performed significantly well, achieving an accuracy of 90%. This provides an effective approach for predicting bond defaults and offers valuable insights for investors in assessing bond default risks.

Keywords

Bond Defaults, Financial Indicators, LDA, SVM

1. 引言

自从 2014 年我国债券市场发生第一起债券违约，在随后的几年里，我国债券市场违约状况频发，相对于 2014 年以前债券市场的刚性兑付的情况，债券违约情况的出现是一个市场健康发展应有的状况。截至目前，我国债券市场陆续出现了超过了上百支违约样本。对债券市场违约风险进行研究是很有必要的[1]-[3]。

公司债券发生违约是受到宏观经济、企业经营状况、行业水平等多方面影响。然而，宏观经济受经济周期影响，行业水平受到当时的宏观政策等方面影响。这些部分影响会长期影响所有企业的平均水平。在相对稳定的宏观经济下，企业自身经营状况才是对债券违约起决定性的作用。公司经营情况反应在每年的公司财务指标上，通过对公司的一些财务指标的研究，如流动比率、速动比率、财务杠杆等指标的变化，能进一步发现公司的财务状况，进一步分析公司违约债券的可能性。通过分析各个因素的作用，这对投资者提前预警，以及企业在经营中生产调整有一定的指导意义。

2. 相关研究

自从债券市场出现违约之后，出现大量学者对其违约预测进行研究。Jin-Chuan Dua 等人提出一种前向强度模型，分析几个常用因素与公司特有属性对违约预测的重要性[4]。张荀杨(2017)对因变量划分，即以违约情况划分为不同类别，以多元回归的方式构建多元逻辑回归模型进行实证研究[5]。王秋龙(2018)通过选取 12 个指标 2 个非财务变量构建两个时间点模型，利用 logit 回归建模得到越靠近违约时间点，流动现金等因素与企业违约有较高的相关性，越靠前时间点资产结构影响更大[6]。魏国健(2018)通过市场违约数据，结合 KMV (Kohn-Merchant-Vasicek, KMV)模型和 logit 模型构建混合模型进行实证分析，结果显示混合模型与 logit 模型一样具有较高的准确性，能较好地识别违约风险[7]。关然(2019)通过吸收 Z-score 模型财务指标，KMV 模型的违约距离等，利用 logit 模型构建出更适用的公司违约风险量化模型[8]。曹昱(2019)通过构建 BP-KMV 模型对上市民营企业数据进行实证分析，通过因子分析法降低变量维度提高模型准确性[9]。上述研究考虑逻辑回归并引入 KMV 模型结果。

随着研究深入，对于预测问题更多关注准确性。因此，最近研究会在基础逻辑回归模型下引入集成学习、深度学习等。吴育辉等选取财务指标与非财务指标，搭建了基于机器学习算法 SMOTETomek-GWO-XGBoost 的债券违约风险预警模型。其有较高的准确性、召回率等[10]。陈湘州等将逻辑回归、随机森林等算法融合，提出了 LR-RF-XGBoost 债券违约预警模型对违约进行预警研究[11]。Chenxiang Zhang 等发现引入卷积神经网络模型对企业违约预测中的表现相较于随机森林等效果更佳[12]。

然而上述研究中，针对债券违约问题，一方面是利用传统经济理论，通过引入 KMV 模型，结合逻辑回归模型对债券进行违约预测，其具有较强的解释性。但是现实中可能并非完全满足模型假设条件。另

一方面从机器学习出发，通过引入集成学习如 XGBOOST 模型、深度学习模型对违约进行预测，在不考虑经济意义条件下，该方法更适用。

对于目前债券市场而言，违约样本毕竟少数。从周期的财务指标分析公司，所涉及维度较高。相对于其他二分类模型而言，SVM 有着较大的优势。本文采用 SVM (Support Vector Machine, SVM)支持向量机模型进行研究，选择 150 家公司作为研究对象，选用数据是公司一年四个季度财务数据。由于数据存在高相关性因此采用 LDA 方法进行降维。构建 LDA-SVM 模型进行建模研究，并与多个二分类模型效果进行比较。

3. 模型建立与评估方法

3.1. LDA-SVM 模型

SVM 是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器。相对于其他二分类算法模型对比，如二分类线性模型逻辑回归，以及神经网络模型，SVM 有明确的几何解释，通过定义边界二分类最大化边界间隔有助于直观理解分类决策同时具有较强的泛化能力。相较于 xgboost 等决策树类模型，支持向量机能够处理相对于样本量维度数据略高的数据，避免决策树表现不稳定的情况。此外，在处理非平衡数据方面，SVM 可以调整权重值等应对，而逻辑回归、决策树等对非平衡数据处理的表现较差。本文以公司财务指标分析公司各项抵御风险能力所涉及的指标数量较多，即使是本文所选择财务指标共计 15 个，同时时间跨度 4 个季度，变量维度共计 60 个，而样本量为 150 个样本，同时因为违约企业与非违约企业样本数存在不平衡性。因此，采用 SVM 模型作为核心预测模型。

对于 SVM 有如下相关理论。

假设一个特征空间上的训练数据集(线性可分)：

$$T=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_N,y_N\right)\right\} \tag{1}$$

其中，$x_i\in\mathbb{R}^n$，$y_i\in\left\{-1,1\right\}$，$i=1,2,\cdots,N$，$x_i$ 为第 i 个特征向量，y_i 为类标记，当为正例是值为 1，反之则为−1。

设存在数据集 T 和超平面 $\boldsymbol{\omega}\cdot\boldsymbol{x}+b=0$，定义超平面关于样本点 $\left(x_i,y_i\right)$ 的几何间隔为：

$$\gamma_i=y_i\left(\frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|}\cdot\boldsymbol{x_i}+\frac{b}{\|\boldsymbol{\omega}\|}\right) \tag{2}$$

超平面关于所有样本点的几何间隔的最小值为：

$$\gamma=\min_{i=1,2,\cdots,N}\gamma_i \tag{3}$$

因此，SVM 模型求解可以变成线性优化问题，即能够满足下面方程组的最优解就是最优线性分类器模型参数。

$$\begin{cases}\max\limits_{\boldsymbol{\omega},b}\gamma_i\\ \text{s.t. } y_i\left(\dfrac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|}\cdot\boldsymbol{x}_i+\dfrac{b}{\|\boldsymbol{\omega}\|}\right)>\gamma,i=1,2,\cdots,N\end{cases} \tag{4}$$

对于公司多周期财务指标除了同一时期不同指标之间存在相关性之外，其相近周期相同指标之间也存在强相关性，若只依靠 SVM 模型核函数映射划分难以寻找合适的映射空间。因此可以考虑先将数据进行映射降维，提升划分的准确性，再利用 SVM 模型进行预测。相较于传统的 PCA 降维方式，这里选用线性判别分析(LDA, Linear Discriminant Analysis)方式，其更加适用分类问题。LDA 目标是找到一个新的特征空间，使得在这个空间中，不同类别的数据点尽可能地分开，同时同一类别的数据点尽可能地聚集在一起。其基本原理如下：

假设数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，定义 $N_j(j=0,1)$ 为第 j 类样本个数，X_j 为第 j 类样本的集合，而 $\mu_j(j=0,1)$ 为第 j 类样本的均值向量，定义 $\Sigma_j(j=0,1)$ 为第 j 类样本的协方差矩阵。则

$$\mu_j=\frac{1}{N_j}\sum_{x\in X_j}x(j=0,1) \tag{5}$$

$$\Sigma_j=\sum_{x\in X_j}(x-\mu_j)(x-\mu_j)^{\mathrm{T}},j=0,1 \tag{6}$$

定义类内散度矩阵 S_W 为：

$$S_W=\sum_{j=0,1}\Sigma_j=\sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^{\mathrm{T}}+\sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^{\mathrm{T}} \tag{7}$$

定义类间散度矩阵 S_b 为：

$$S_b=(u_0-u_1)(u_0-u_1)^{\mathrm{T}} \tag{8}$$

对于二分类任务，需要将数据投射到直线上，令投射直线向量为 $\boldsymbol{w}$，则对于任意样本 x_i，其直线 $\boldsymbol{w}$ 的投影为 $\boldsymbol{w}^{\mathrm{T}}x_i$，对于两类别中心点 u_0,u_1，在直线 $\boldsymbol{w}$ 的投影为 $\boldsymbol{w}^{\mathrm{T}}u_0,\boldsymbol{w}^{\mathrm{T}}u_1$。使得映射后不同类别的数据的类别中心之间的距离 $\left\|\boldsymbol{w}^{\mathrm{T}}u_0-\boldsymbol{w}^{\mathrm{T}}u_1\right\|^2$ 尽可能大，同时在同一类别中的样本点尽可能的近，即 $\boldsymbol{w}^{\mathrm{T}}\Sigma_0\boldsymbol{w},\boldsymbol{w}^{\mathrm{T}}\Sigma_1\boldsymbol{w}$ 尽可能小。因此最优化目标为：

$$\arg\max J(w)=\frac{\left\|\boldsymbol{w}^{\mathrm{T}}u_0-\boldsymbol{w}^{\mathrm{T}}u_1\right\|^2}{\boldsymbol{w}^{\mathrm{T}}\Sigma_0\boldsymbol{w}+\boldsymbol{w}^{\mathrm{T}}\Sigma_1\boldsymbol{w}}=\frac{\boldsymbol{w}^{\mathrm{T}}S_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}S_W\boldsymbol{w}} \tag{9}$$

该优化目标实质上是一个瑞利商，根据瑞利商的性质得到最优的投射直线 $\boldsymbol{w}'$。

因此 LDA-SVM 模型是将 LDA 降维方法与 SVM 相结合而成。其中 LDA 按照之前理论进行降维，再经过 SVM 模型得到预测结果。

3.2. 模型评估方法

债券违约与否是一个二分类问题，因此可以通过混淆矩阵观察模型的效果。其中混淆矩阵结构如下表 1 所示。

Table 1. Confusion matrix example table
表 1. 混淆矩阵例表

	预测为正类	预测为负类
实际为正类	TP	FN
实际为负类	FP	TN

根据混淆矩阵可以计算准确率(Accuracy)、F1-score 指标评价模型的好坏。其中计算方式如下：

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{10}$$

$$\text{F1-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

其中，Precision 与 Recall 计算方式如下：

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

同时，通过混淆矩阵衍生出来的假正率(FPR)、召回率(TPR)两个指标进一步绘制 ROC 曲线()。通过计算 AUC 面积(曲线右下部分围成面积)评价模型分类效果的好坏。其中 FPR、TPR 的计算公式如下：

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{14}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

4. 实验分析

4.1. 数据来源与数据处理

本文选用数据是 150 家发行公司债的公司，其中涉及违约债券 34 家，非违约债券 116 家，涉及行业不固定。存在样本量的不平衡性。这是符合债券市场现状的，打破刚性兑付的债券市场上，非违约企业仍然占据大多数。选取相关指标根据公司盈利能力、偿债能力、成长能力等方面考虑共计 15 个指标，具体变量如下表 2 所示。选取时间距离发生违约时间点一年的财务数据，最终变量合计 60 个维度。因此，样本数据特征是具有不平衡性、以及维度较高的特点，但这是符合债券市场现状，总体违约量少，各项评价指标数目较多。

Table 2. Indicators-variable
表 2. 指标－变量

指标	变量名
违约与否	Y
营业总收入(同比增长率)	II
总负债(同比增长率)	ADI
净利润(同比增长率)	NPI
销售毛利率	GM
净资产收益率 ROE (平均)	ROE
资产负债率	ALR
流动负债/负债合计	AL2L
流动比率	CR
速动比率	QR
应收账款周转率	ATR
应付账款周转率	APR
带息债务	IBD
净利润(TTM)	NP
营业总收入(TTM)	II.1
营业总成本(TTM)	AC

在数据集划分方面，本文将 150 个样本数据按照 8:2 比例随机划分为训练集与测试集。因为违约样本为 34 条，同时为了保证每个集合中都包含违约样本，因此采用分层抽样的方式进行划分，即违约样本中随机等可能抽取 27 个样本与非违约样本中等可能随机抽取 93 个样本共同组成训练集，其余样本为测试集。

4.2. 实验与分析

将训练集样本带入模型中进行训练，同时将训练完成的模型通过测试集验证。其结果如下表 3 所示。

Table 3. Experimental results confusion matrix
表 3. 实验结果混淆矩阵

	预测不违约	预测违约
实际不违约	23	1
实际违约	2	4

通过混淆矩阵计算出模型的准确率为 90%，f1 得分为 90%。得到 ROC 曲线如下图 1 所示：

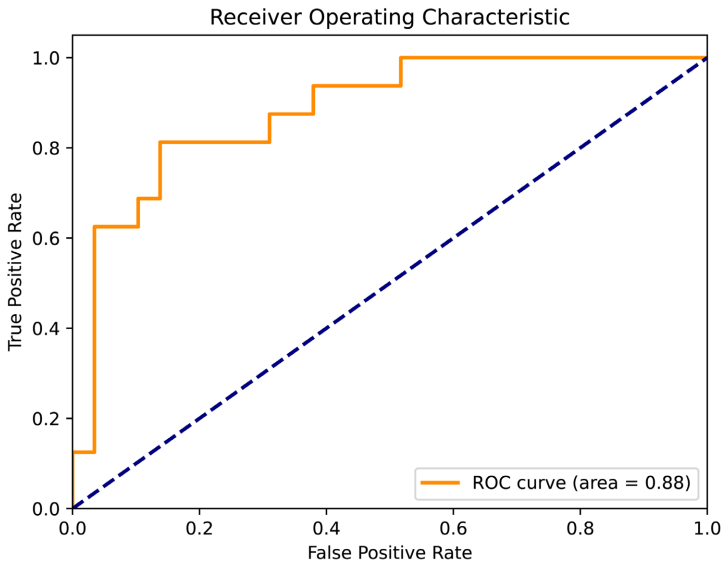

Figure 1. The ROC curves of LDA-SVM
图 1. LDA-SVM 模型 ROC 曲线

从图 1 中可以看到模型效果显著，AUC 面积为 0.99，说明模型预测效果好。
接着与二分类常规模型进行对比，以准确率作为评价指标。其结果如下表 4 所示。

Table 4. Individual model results
表 4. 各个模型结果

模型	准确率 acc
k 近邻	0.6
决策树	0.73
svm	0.73

续表

logistic	0.64
随机森林	0.76
自增强算法	0.76
Lightboost	0.78
Xgboost	0.8
LDA-SVM	0.9

ROC 曲线如下图 2、图 3 所示：

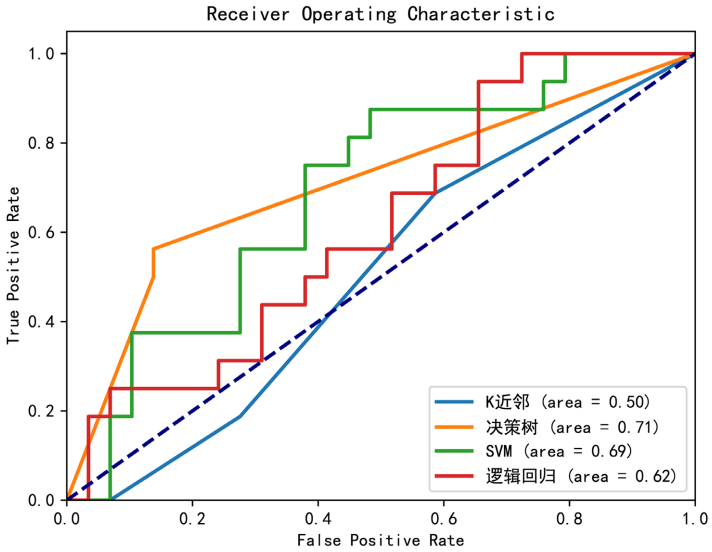

Figure 2. ROC curves of partial models (1)
图 2. 部分模型 ROC 曲线(一)

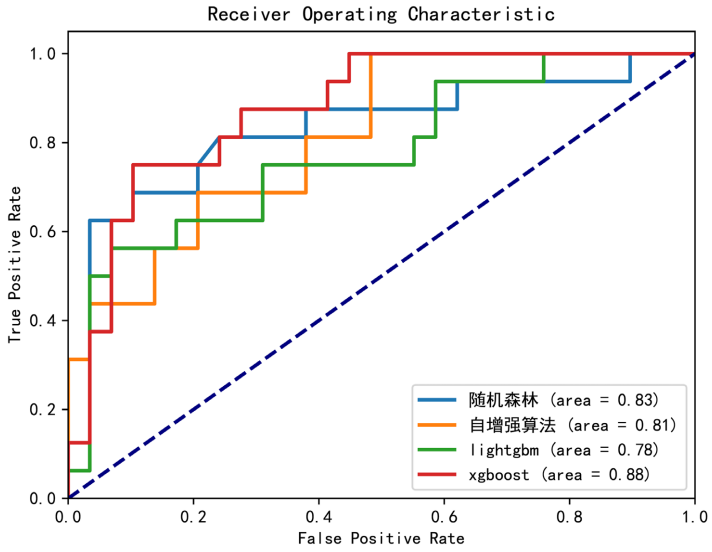

Figure 3. ROC curves of partial models (2)
图 3. 部分模型 ROC 曲线(二)

从表 4 中可知，常规分类模型如 k 近邻等对违约预测效果不佳，准确率在 0.8 以下。其次是集成学习算法例如 xgboost 算法对预测效果有明显提高但是也只是达到 0.8。通过对比，本文提出的 LDA-SVM 模型效果显著。能更有效预测债券违约风险。

5. 结论与展望

本文以债券市场中 150 家公司作为研究对象，选取企业盈利能力、偿债能力、成长能力相关财务指标作为对公司债券是否违约的潜在因素分析。通过选取一年共四个季度的数据进行分析。考虑各个变量之间存在相关性问题，为提高预测精度，提出了 LDA-SVM 模型对债券是否违约进行建模分析。实验结果表明，该模型能显著地预测债券违约情况。同时再比较多个二分类模型，结果表明本文所提出模型效果显著，准确率达到 90%，能更精确对债券违约与否做出判别。

本文是根据一年期数据分析未来债券违约。未来对债券违约的研究，还需要考虑债券市场的走向，债券价格从一定程度上反应背后的企业经营情况。考虑引入经典的 KMV 模型，虽然存在公司是非上市公司，难以通过 KMV 模型得到违约距离，但可以进一步分析上市公司的违约风险。

参考文献

[1] 吴秋余. 7 月我国债券市场发行超 6.6 万亿元[N]. 人民日报, 2024-08-28(010).

[2] 乔君, 白俊, 袁勋. 债券市场“刚性兑付”打破与企业自愿性业绩预告[J/OL]. 财贸研究, 2024: 1-20. http://kns.cnki.net/kcms/detail/34.1093.F.20240506.1851.002.html, 2024-08-29.

[3] 李丹凤. 我国债券市场研究综述与展望——基于 CiteSpace 的可视化分析[J]. 中国商论, 2024(6): 100-103.

[4] Duan, J.C., Sun, J. and Wang, T. (2012) Multiperiod Corporate Default Prediction—A Forward Intensity Approach. *Journal of Econometrics*, **170**, 191-209. https://doi.org/10.1016/j.jeconom.2012.05.002

[5] 张荀杨. 基于多元 LOGIT 模型的公司债券违约因素实证研究[D]: [硕士学位论文]. 泉州: 华侨大学, 2017.

[6] 王秋龙. 基于 logit 模型的我国信用债市场的信用风险研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2018.

[7] 魏国健. 基于 KMV-LOGIT 混合模型的信用债券违约风险度量与实证研究[D]: [硕士学位论文]. 合肥: 中国科学技术大学, 2018.

[8] 关然. 基于 Logit 模型的上市公司违约风险量化评估[D]: [硕士学位论文]. 广州: 对外经济贸易大学, 2019.

[9] 曹昱. 基于 BP-KMV 组合模型的民营企业信用风险度量实证研究[D]: [硕士学位论文]. 济南: 山东大学, 2019.

[10] 吴育辉, 刘忻忻, 陈韫妍. 债券违约预警模型的优化与提升——基于 SMOTETomek-GWO-XGBoost 的方法[J]. 会计之友, 2024(6): 73-81.

[11] 陈湘州, 刘佳. 基于 LR-RF-XGBoost 的债券违约风险预警[J]. 湖南科技大学学报(自然科学版), 2024, 39(1): 115-124.

[12] Zhang, C., Zhang, F., Chen, N., *et al*. (2022) Application of Artificial Intelligence Technology in Financial Data Inspection and Manufacturing Bond Default Prediction in Small and Medium-Sized Enterprises (SMEs). *Operations Management Research*, **15**, 941-952.